Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

Mon. Not. R. Astron. Soc. **409,** 355–370 (2010) doi:10.1111/j.1365-2966.2010.17313.x

Bayesian non-linear large-scale structure inference of the Sloan Digital Sky Survey Data Release 7

Jens Jasche,[1⋆] Francisco S. Kitaura,[2] Cheng Li[1] and Torsten A. Enßlin[1]

[1]*Max-Planck-Institut für Astrophysik, Karl-Schwarzschild Straße 1, D-85748 Garching, Germany*
[2]*SNS, Scuola Normale Superiore, Piazza dei Cavalieri 7, 56126 Pisa, Italy*

Accepted 2010 July 6. Received 2010 June 23; in original form 2009 November 12

ABSTRACT

In this work, we present the first non-linear, non-Gaussian full Bayesian large-scale structure analysis of the cosmic density field conducted so far. The density inference is based on the Sloan Digital Sky Survey (SDSS) Data Release 7, which covers the northern galactic cap. We employ a novel Bayesian sampling algorithm, which enables us to explore the extremely high dimensional non-Gaussian, non-linear lognormal Poissonian posterior of the three-dimensional density field conditional on the data. These techniques are efficiently implemented in the Hamiltonian Density Estimation and Sampling (HADES) computer algorithm and permit the precise recovery of poorly sampled objects and non-linear density fields. The non-linear density inference is performed on a 750-Mpc cube with roughly 3-Mpc grid resolution, while accounting for systematic effects, introduced by survey geometry and selection function of the SDSS, and the correct treatment of a Poissonian shot noise contribution. Our high-resolution results represent remarkably well the cosmic web structure of the cosmic density field. Filaments, voids and clusters are clearly visible. Further, we also conduct a dynamical web classification and estimate the web-type posterior distribution conditional on the SDSS data.

Key words: methods: data analysis – methods: numerical – cosmology: observations – large-scale structure of Universe.

1 INTRODUCTION

Observations of the large-scale structure have always attracted enormous interest, since they contain a wealth of information on the origin and evolution of our Universe. The details of structure formation are very complicated and involve many different physical disciplines ranging from quantum field theory, general relativity or modified gravity to the dynamics of collisionless matter and the behaviour of the baryonic sector. Throughout cosmic history, the interplay of these different physical phenomena has therefore left its imprints in the matter distribution surrounding us. Probes of the large-scale structure, such as large galaxy surveys, hence enable us to test current physical and cosmological theories and will generally further our understanding of the Universe.

Especially a cosmographical description of the matter distribution will permit us to study details of structure formation mechanisms and the clustering behaviour of galaxies as well as it will provide information on the initial fluctuations and large-scale cosmic flows. For this reason, several different methods to recover the three-dimensional density or velocity field from galaxy observations have been developed and applied to existing galaxy surveys (Ebeling & Wiedenmann 1993; Hoffman 1994; Lahav 1994; Lahav et al. 1994; Fisher et al. 1995; Zaninetti 1995; Zaroubi et al. 1995; Webster, Lahav & Fisher 1997; Zaroubi, Hoffman & Dekel 1999; van de Weygaert & Schaap 2001; Erdoğdu et al. 2006, 2004; Kitaura, Jasche & Metcalf 2010). In particular, recently Kitaura et al. (2009) presented a high-resolution three-dimensional Wiener reconstruction of the Sloan Digital Sky Survey (SDSS) Data Release 6 data, which demonstrated the feasibility of high precision density field inference from galaxy redshift surveys. These three-dimensional density maps are interesting for a variety of different scientific applications, such as studying the dependence of galaxy properties on their cosmic environment, increasing the detectability of the integrated Sachs–Wolfe effect in the cosmic microwave background or performing constrained simulations (see e.g. Bistolas & Hoffman 1998; Klypin et al. 2003; Frommert, Enßlin & Kitaura 2008; Lee & Lee 2008; Lee & Li 2008; Libeskind et al. 2010; Martinez-Vaquero et al. 2009).

However, modern precision cosmology demands an increasing control of observational systematic and statistical uncertainties and the means to propagate them to any finally inferred quantity in order not to draw wrong conclusion on the theoretical model to be tested. For this reason, here we present the first application of the new

⋆E-mail: jjasche@mpa-garching.mpg.de

Bayesian large-scale structure inference computer algorithm Hamiltonian Density Estimation and Sampling (HADES) to data (see Jasche & Kitaura 2010, for a description of the algorithm). HADES performs a full scale non-linear, non-Gaussian Markov chain Monte Carlo analysis by drawing samples from the lognormal Poissonian posterior of the three-dimensional density field conditional on the data. This extremely high dimensional posterior distribution, consisting of $\sim 10^6$ or more free parameters, is explored via a numerically efficient Hamiltonian sampling scheme which suppresses the random walk behaviour of conventional Metropolis Hastings algorithms by following persistent trajectories through the parameter space (Duane et al. 1987; Neal 1993, 1996). The advantages of this method are manyfold. Beside correcting observational systematics introduced by survey geometry and selection effects, the exact treatment of the non-Gaussian behaviour and structure of the Poissonian shot noise contribution of discrete galaxy distributions permits very accurate recovery of poorly sampled objects, such as voids and filaments. In addition, the lognormal prior has been demonstrated to be an adequate statistical description for the present density field and hence enables us to infer the cosmic density field deep into the non-linear regime (see e.g. Hubble 1934; Peebles 1980; Coles & Jones 1991; Gaztanaga & Yokoyama 1993; Kayo, Taruya & Suto 2001). The important thing to remark about HADES is that it does not only yield a single estimate, such as a mean, mode or variance, in fact it provides a sampled representation of the full non-Gaussian density posterior. This posterior encodes the full non-linear and non-Gaussian observational uncertainties, which can easily be propagated to any finally inferred quantity.

The application of HADES to SDSS data is the first non-linear, non-Gaussian full Bayesian large-scale structure analysis conducted so far (SDSS; York et al. 2000). In particular, we applied our method to the recent SDSS Data Release 7 (DR7) (Abazajian et al. 2009) and produced about 3 terabyte (TB) of valuable scientific information in the form of 40 000 high-resolution non-linear density samples. The density inference is conducted on an equidistant cubic grid with a side length of 750 Mpc consisting of 256^3 volume elements. The recovered density field clearly reveals the cosmic web structure, consisting of voids, filaments and clusters, of the large-scale structure surrounding us.

These results provide the basis for forthcoming investigations on the clustering behaviour of galaxies in relation to their large-scale environment. Such analyses yield valuable information about the formation and evolution of galaxies. For example, it has been known since long that physical properties such as morphological type, colour, luminosity, spin parameter, star formation rate, concentration parameter, etc., are functions of the cosmic environment (see e.g. Dressler 1980; Postman & Geller 1984; Whitmore, Gilmore & Jones 1993; Lewis et al. 2002; Gómez et al. 2003; Goto et al. 2003; Blanton et al. 2005; Kuehn & Ryden 2005; Rojas et al. 2005; Bernardi et al. 2006; Choi, Park & Vogeley 2007; Park et al. 2007; Lee & Lee 2008; Lee & Li 2008).

In this work, we conduct a preliminary examination of the dependence of stellar mass $M_\star$ and $g - r$ colour of galaxies on their large-scale environment. However, more thorough investigations will be presented in following works. Analysing galaxy properties in the large-scale environment also requires the classification of the large-scale structure into different cosmic web types. We do so by following the dynamic cosmic web-type classification procedure proposed by Hahn et al. (2007) with the extension proposed by Forero-Romero et al. (2009). The application of this procedure to our results yields the cosmic-web-type posterior, which provides the probability of finding a certain web type (void, sheet, filament, halo) at a given position in the volume conditional on the SDSS data. This permits simple propagation of all observational uncertainties to the final analysis of galaxy properties.

This paper is structured as follows. We start by a brief review of the methodology in Section 2, particularly describing the lognormal Poissonian posterior and the Bayesian computer algorithm HADES. Additionally, here we describe the dynamic web classification procedure as mentioned above. In Section 3, we give a description of the SDSS DR7 data and present necessary data preparation steps required to apply the analysis procedure. Specifically, we describe the preparation of the linear observation response operator and the creation of the three-dimensional data cube. In Section 4, we present the results obtained from the non-linear, non-Gaussian sampling procedure. We start by analysing the convergence behaviour of the Markov chain via a Gelman & Rubin diagnostic, followed by a discussion of the properties of individual Hamiltonian samples. Here we also provide estimates for the ensemble mean density field and corresponding variance. These fields depict remarkably well the properties of the cosmic web consisting of voids, filaments and haloes. Based on these results, we perform a simple cosmic web classification in Section 5. In Section 6, we present a preliminary examination on the correlation between the large-scale environment of galaxies and their physical properties. In particular, here we study the stellar mass and $g - r$ colour of galaxies in relation to the density contrast δ. We conclude the paper in Section 7 by summarizing and discussing the results.

2 METHODOLOGY

In this section, we give a brief review of the methods used for the large-scale structure inference. In particular, we discuss the lognormal Poissonian posterior and the according data model. Further, we give a description of the HADES algorithm and a dynamic cosmic web classification procedure.

2.1 Lognormal Poissonian posterior

Precision inference of the large-scale structure in the mildly and strongly non-linear regime requires detailed treatment of the non-Gaussian behaviour of the large-scale structure posterior. Although the exact probability distribution for the density field in these regimes is not known, however it has been already suggested for a long time that the fully evolved non-linear matter field can be well described by lognormal statistics (see e.g. Hubble 1934; Peebles 1980; Coles & Jones 1991; Gaztanaga & Yokoyama 1993; Kayo et al. 2001). This phenomenological guess has been justified by the theoretical considerations of Coles & Jones (1991). They argue that assuming Gaussian initial conditions in the density and velocity distributions will lead to a lognormally distributed density field. It is a direct consequence of the continuity equation or the conservation of mass. In addition, the validity of the lognormal distribution as a description of the statistical properties of non-linear density fields has been evaluated in Kayo et al. (2001). In this work, they studied the probability distribution of cosmological non-linear density fluctuations from N-body simulations with Gaussian initial conditions. They found that the lognormal distribution accurately describes the non-linear density field even up to values of the density contrast of $\delta \sim 100$. In addition, recently, Kitaura et al. (2009) analysed the statistical properties of the SDSS DR6 Wiener-reconstructed density field and confirmed a lognormal behaviour.

For all these reasons, we believe that the statistical behaviour of the non-linear density field can be well described by a multivariate

lognormal distribution, as given by

$$\mathcal{P}(\{s_k\}|Q) = \frac{1}{\sqrt{2\pi\det(Q)}} \mathrm{e}^{-\frac{1}{2}\sum_{ij}(ln(1+s_i)+\mu_i)Q_{ij}^{-1}(ln(1+s_j)+\mu_j)} \times \prod_k \frac{1}{1+s_k}\,, \quad (1)$$

where s_i is the density signal at the three-dimensional Cartesian position $\boldsymbol{x}_i$, Q is the covariance matrix of the lognormal distribution and μ_i describes a constant mean field given by

$$\mu_i = \frac{1}{2}\sum_{l,m} Q_{lm}\,. \quad (2)$$

This probability distribution seems to be an adequate prior choice for reconstructing the present density field.

Studying the actual matter distribution of the Universe requires us to draw inference from some observable tracer particle, such as a set of observed galaxies. Assuming galaxies to be discrete particles, their distribution can be described as a specific realization drawn from an inhomogeneous Poisson process (see e.g. Layzer 1956; Peebles 1980; Martínez & Saar 2002). The according probability distribution is given as

$$\mathcal{P}\left(\left\{N_k^{\mathrm{g}}\right\}|\{\lambda_k\}\right) = \prod_k \frac{(\lambda_k)^{N_k^{\mathrm{g}}}\mathrm{e}^{-\lambda_k}}{N_k^{\mathrm{g}}!}\,, \quad (3)$$

where N_k^{g} is the observed galaxy number at position $\boldsymbol{x}_k$ in the sky and λ_k is the expected number of galaxies at this position. The mean galaxy number is related to the signal s_k via

$$\lambda_k = R_k\bar{N}(1+B(s)_k)\,, \quad (4)$$

where R_k is a linear response operator, incorporating survey geometries and selection effects; $\bar{N}$ is the mean number of galaxies in the volume and $B(x)_k$ is a non-linear, non-local, bias operator at position $\boldsymbol{x}_k$. The lognormal prior given in equation (1) together with the Poissonian likelihood given in equation (3) yields the lognormal Poissonian posterior, for the density contrast s_k given some galaxy observations N_k^{g}:

$$\mathcal{P}\left(\{s_k\}|\left\{N_k^{\mathrm{g}}\right\}\right) = \frac{\mathrm{e}^{-\frac{1}{2}\sum_{ij}(ln(1+s_i)+\mu_i)Q_{ij}^{-1}(ln(1+s_j)+\mu_j)}}{\sqrt{2\pi\det(Q)}} \prod_l \frac{1}{1+s_l} \times \prod_k \frac{(R_k\bar{N}(1+B(s)_k))^{N_k^{\mathrm{g}}}\mathrm{e}^{-R_k\bar{N}(1+B(s)_k)}}{N_k^{\mathrm{g}}!}. \quad (5)$$

It is important to note that this is a highly non-Gaussian distribution, and non-linear reconstruction methods are required in order to perform accurate matter field reconstructions in the non-linear regime. For example, estimating the maximum a posteriori values from the lognormal Poissonian distribution involves the solution of implicit equations. Several attempts to use a lognormal Poissonian posterior for density inference have been presented in the literature. These attempts date back to Sheth (1995) who proposed to use a variable transformation in order to derive a generalized Wiener filter for the lognormal distribution. This approach, however, yielded a very complex form for the noise covariance matrix making applications to real data sets impractical. The first successful application of the lognormal Poissonian distribution for density inference was presented by Saunders et al. (2000). Their method is based on the expansion of the density logarithm into spherical harmonics (Saunders & Ballinger 2000). More accurate schemes based on maximum and mean posteriori principles were derived by Enßlin, Frommert & Kitaura (2009). Recently, an implementation of the maximum a posteriori scheme was presented and thoroughly tested by Kitaura et al. (2010). They found that, assuming a linear bias, the lognormal Poissonian posterior permits recovery of the density field deep in the non-linear regime up to values $\delta \geq 1000$ of the density contrast. Finally, Jasche & Kitaura (2010) developed the Hamiltonian density estimation and sampling scheme to map out the posterior probability distribution.

2.2 HADES

As already described above, the Bayesian non-linear large-scale structure inference requires sampling from non-Gaussian posterior distributions. In order to do so, we developed HADES (see Jasche et al. 2010, for more details). HADES explores the very high dimensional parameter space of the three-dimensional density field via a Hamiltonian Monte Carlo (HMC) sampling scheme. Unlike conventional Metropolis Hastings algorithms, which move through the parameter space by a random walk and therefore require a prohibitive number of steps to explore high dimensional spaces, the HMC sampler suppresses random walk behaviour by introducing a persistent motion of the Markov chain through the parameter space (Duane et al. 1987; Neal 1993, 1996). In this fashion, the HMC sampler maintains a reasonable efficiency even for high dimensional problems (Hanson 2001). Since it is a fully Bayesian method, the scientific output is not a single estimate but a sampled representation of the multidimensional lognormal Poissonian posterior distribution given in equation (5). Given this representation of the posterior any desired statistical summary, such as mean, mode or variances, can easily be calculated. Further, any uncertainty can seamlessly be propagated to the finally estimated quantities, by simply applying the according estimation procedure to all Hamiltonian samples. For a detailed description of the theory behind the large-scale structure sampler and its numerical implementation, see Jasche et al. (2010).

2.3 Classification of the cosmic web

The results generated by the Hamiltonian sampler HADES will permit a variety of scientific analyses of the large-scale structure in the observed Universe. An interesting example is to classify the cosmic web, in particular identifying different types of structures in the density field. Such an analysis, for example, is valuable for studying the environmental dependence of galaxy formation and evolution (see e.g. Lee & Lee 2008; Lee & Li 2008). Since the structure classification is not always unique, we provide the full Bayesian posterior distribution of the structure type at a given position conditional on the observations.

However, to do so we first need a means to identify different structure types from the density field. Numerous methods for structure analysis have been presented in the literature (see e.g. Lemson & Kauffmann 1999; Colberg et al. 2005, 2008; Novikov, Colombi & Doré 2006; Aragón-Calvo et al. 2007; Hahn et al. 2007; Forero-Romero et al. 2009). In principle, all of these methods can be applied for the analysis of the Hamiltonian samples; however, for the purpose of this paper, we follow the dynamical cosmic web classification procedure proposed by Hahn et al. (2007). They propose to classify the large-scale structure environment into four web types (voids, sheets, filaments and haloes) based on a local-stability criterion for the orbits of test particles. The basic idea of this dynamical classification approach is that the eigenvalues of the deformation tensor characterize the geometrical properties of each point in space. The deformation tensor T_{ij} is given by the Hessian of the

Table 1. Rules for the dynamic classification of web types.

Structure type	Rule
Void	$\lambda_1, \lambda_2, \lambda_3 < \lambda_{\rm th}$
Sheet	$\lambda_1 > \lambda_{\rm th}$ and $\lambda_2, \lambda_3 < \lambda_{\rm th}$
Filament	$\lambda_1, \lambda_2 > \lambda_{\rm th}$ and $\lambda_3 < \lambda_{\rm th}$
Halo	$\lambda_1, \lambda_2, \lambda_3 > \lambda_{\rm th}$

gravitational potential Φ:

$$T_{ij} = \frac{\partial^2 \Phi}{\partial x_i \, \partial x_j} \,, \tag{6}$$

with Φ being the rescaled gravitational potential given as (see Forero-Romero et al. 2009)

$$\nabla^2 \Phi = \delta \,. \tag{7}$$

It is important to note that the deformation tensor and the rescaled gravitational potential are both physical quantities, and hence their calculation requires the availability of a full physical density field in contrast to a smoothed mean reconstruction of the density field. As was already mentioned above, and will be clarified in Section 4.2, the Hamiltonian samples provide such required full physical density fields. The deformation tensor can therefore be easily calculated for each Hamiltonian sample from the Fourier space representation of equation (6). Each spatial point can then be classified as a specific web type by considering the three eigenvalues, $\lambda_1 \geq \lambda_2 \geq \lambda_3$, of the deformation tensor. Namely, a void point corresponds to no positive eigenvalue, a sheet to one, a filament to two and a halo to three positive eigenvalues (Forero-Romero et al. 2009). The interpretation of this rule is straightforward, as the sign of a given eigenvalue at a given position defines whether the gravitational force at the direction of the principal direction of the corresponding eigenvector is contracting (positive eigenvalues) or expanding (negative eigenvalues). However, Forero-Romero et al. (2009) found that rather than using a threshold value $\lambda_{\rm th}$ of zero, different positive values can yield better web classifications. For this reason, in this work, we use the extended classification procedure as proposed by Forero-Romero et al. (2009). The structures are then classified according to the rules given in Table 1. By applying this classification procedure to all Hamiltonian samples, we are able to estimate the web-type posterior $\mathcal{P}(\{\mathrm{T}_i(\boldsymbol{x}_k)\}|\{N_k^{\rm g}\}, \lambda_{\rm th})$ of four different web types [$T_1(\boldsymbol{x}_k) =$ void, $T_2(\boldsymbol{x}_k) =$ sheet, $T_3(\boldsymbol{x}_k) =$ filament, $T_4(\boldsymbol{x}_k) =$ halo] conditional on the observations and the threshold criterion $\lambda_{\rm th}$.

3 DATA

In this section, we describe the SDSS galaxy sample used for the analysis. Additionally, we discuss the data preparation steps required to perform the three-dimensional density inference procedure.

3.1 The SDSS galaxy sample

We use data from sample DR72 of the New York University Value Added Catalogue (NYU-VAGC).[1] This is an update of the catalogue constructed by Blanton et al. (2005) and is based on the final data release DR7 (Abazajian et al. 2009) of the SDSS (York et al. 2000). Starting from sample DR72, we construct a magnitude-limited sample of galaxies with spectroscopically measured redshifts in the range $0.001 < z < 0.4$, r-band Petrosian apparent magnitude $r \leq 17.6$ after correction for Galactic extinction and r-band absolute magnitude $-23 < M_{0.1_r} < -17$. Here $M_{0.1_r}$ is corrected to its $z = 0.1$ value using the K-correction code of Blanton et al. (2003b) and Blanton & Roweis (2007) and the luminosity evolution model of Blanton et al. (2003a). The apparent magnitude limit is chosen in order to get a sample that is uniform and complete over the entire area of the survey. We also restrict ourselves to galaxies located in the main contiguous area of the survey in the northern Galactic cap, excluding the three survey strips in the southern cap (about 10 per cent of the full survey area). In addition, we consider only galaxies which are inside a comoving cube of a side length of 750 Mpc. These restrictions result in a final sample of 463 230 galaxies.

The NYU-VAGC also provides the necessary information to correct for incompleteness in our spectroscopic sample. This includes two parts: a mask which shows which areas of the sky have been targeted and which have not, and a radial selection function which gives the fraction of galaxies in the absolute magnitude range being considered that are within the apparent magnitude range of the sample at a given redshift. The mask defines the effective area of the survey on the sky, which is 6437 deg^2 for the sample we use here. This survey area is divided into a large number of smaller subareas, called *polygons*, for each of which the NYU-VAGC lists a spectroscopic completeness, defined as the fraction of photometrically identified target galaxies in the polygon for which usable spectra were obtained. Over our sample, the average completeness is 0.92.

3.2 Completeness and selection function

Three-dimensional density field inference requires the definition of the linear observation response operator R_k, as given in Section 2.1. This response operator describes to what percentage each volume element of the three-dimensional domain has been observed. It is hence a projection of the product of radial and angular selection function into the three-dimensional voxelized space. In particular, we have to solve the convolution integral:

$$R_k = R(\boldsymbol{x}_k) = \int \mathrm{d}\boldsymbol{y} \; W\left(\boldsymbol{x}_k - \boldsymbol{y}\right) \; f\left(r(\boldsymbol{y})\right) \; M\left(\alpha(\boldsymbol{y}), \delta(\boldsymbol{y})\right) , \tag{8}$$

where $W(\boldsymbol{x})$ is the voxel kernel, $f(r)$ is the radial selection function, with r being the distance from the observer, and $M(\alpha, \delta)$ is the angular selection function, where α and δ are right ascension and declination, respectively. We evaluate this integral numerically for the nearest grid point kernel by following different lines of sight and calculating the contribution of the product of angular and radial selection function to each voxel.

As mentioned above, in this work we used the two-dimensional sky mask and the radial selection function provided by the NYU-VAGC.

3.3 Creating the three-dimensional data cube

The large-scale structure sampler operates on a three-dimensional equidistant grid. In particular, in this work we set up a cubic grid with a side length of 750 Mpc and 256^3 voxels. This amounts to a resolution of $\sim$3−Mpc voxel side length. Since our algorithm relies on the correlation function in comoving space, all calculations are performed with comoving length units rather than with redshift distances. For this reason, we transform all galaxy redshifts z to

[1] http://sdss.physics.nyu.edu/vagc/

comoving distances via the relation

$$r = \int_0^{z_i} \mathrm{d}z \frac{1}{c\, H(z)}, \tag{9}$$

where z_i is the estimated galaxy redshift, c is the speed of light and $H(z)$ is the Hubble parameter given as

$$H(z) = H_0 \sqrt{\Omega_\mathrm{m}\,(1+z)^3 + \Omega_\mathrm{c}\,(1+z)^2 + \Omega_\Lambda}\,. \tag{10}$$

Further, we choose a concordance Λ cold dark matter (ΛCDM) model with a set of cosmological parameters ($\Omega_\mathrm{m} = 0.24$, $\Omega_\mathrm{c} = 0.00$, $\Omega_\Lambda = 0.76$, $h = 0.73$, $H_0 = h\,100\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$; Spergel et al. 2007,). With these definitions, we can calculate the three-dimensional Cartesian coordinates for each galaxy as

$$\begin{aligned} x &= r\,\cos(\delta)\,\cos(\alpha) \\ y &= r\,\cos(\delta)\,\sin(\alpha) \\ z &= r\,\sin(\delta), \end{aligned} \tag{11}$$

where α and δ are the right ascension and declination, respectively. We then sort the galaxy distribution into the three-dimensional equidistant grid via a nearest grid point procedure (see e.g. Hockney & Eastwood 1988). An estimate for the expected number of galaxies $\bar{N}$ can then be calculated as

$$\bar{N} = \frac{\sum_k N_k^\mathrm{g}}{\sum_l R_l} \tag{12}$$

(see e.g. Kitaura et al. 2009; Jasche et al. 2010, for details).

3.4 Physical model

Observations of the galaxy redshifts do not permit direct inference of the underlying matter distribution. Various physical effects such as galaxy biasing and redshift-space distortions must be taken into account for proper analyses. This is of particular relevance for the choice of a power spectrum required for the sampling procedure (see equation 1). However, according to the discussion in Erdoğdu et al. (2004) and Kitaura et al. (2009) these effects can be greatly accounted for in a separate post-processing step, once the continuous expected galaxy density field in redshift space has been obtained. For this reason, here we seek to recover the density field in redshift space permitting us to test various bias models and redshift space distortion correction methods in a subsequent step.

In particular, the relation between the true underlying dark matter density field and the expected continuous galaxy density contrast is generally very complicated and involves non-local and non-linear bias operators. Several non-local bias models have been presented, which mostly aim at correcting the large-scale power in power-spectrum estimation procedures (see e.g. Peacock & Smith 2000; Seljak 2000; Tegmark et al. 2004; Hamann et al. 2008). As described in Sections 2 and 2.2, the Hamiltonian sampler is able to account for such bias models. Note, however, that a specific bias model also fixes the model for the underlying dark matter distribution. Therefore, here, we prefer to follow the approach of previous works of setting the bias operator to a constant linear factor equal to unity (Erdoğdu et al. 2004; Kitaura et al. 2009). In this fashion, one obtains the expected continuous galaxy density contrast. As discussed in Kitaura et al. (2009), the according underlying dark matter distribution can then be simply obtained by deconvolving the results with a specific scale-dependent bias model, permitting us to explore various different bias models.

In a similar manner, one can treat redshift-space distortion effects. These are mainly due to the peculiar velocities of galaxies, which introduce Doppler effects in the redshift measurement (see e.g. Davis & Peebles 1983; Kaiser 1987; Peacock & Dodds 1994; Hamilton 1998). This effect leads to a radial smearing of the observed density field in redshift space and yields elongated structures along the line of sight, the so-called *Finger-of-God* effect.

Additionally, there exists a cosmological redshift-space effect which is sensitive to the global geometry of the Universe. In particular, the comoving separation of a pair of galaxies at $z \gg 0.1$ is not determined only by their observable angular and redshift separations without specifying the geometry or equivalently the matter content of the Universe (Magira, Jing & Suto 2000). This effect yields anisotropies in the matter distribution especially at $z \geq 1$ (see e.g. Alcock & Paczyński 1979; Matsubara & Suto 1996; Ballinger, Peacock & Heavens 1996; Popowski et al. 1998). However, for the volume considered in this work ($z \leq 0.27$), the dominant redshift-space distortions are due to non-linear peculiar motions of galaxies in large overdensities. This effect has pronounced consequences for the power spectrum in redshift space, since it suppresses power on small scales. As demonstrated in Erdoğdu et al. (2004), the redshift-space power spectrum of a fully evolved non-linear matter distribution is very similar to a linear power spectrum at the scales relevant for this work ($k \leq 2\,h\,\mathrm{Mpc^{-1}}$). Here, they used the non-linear power-spectrum fitting formula provided by Smith et al. (2003). However, the exact galaxy power spectrum in redshift space is not known. The work of Tegmark et al. (2006) indicates that the recovered power spectrum of the SDSS main sample is close to a linear power spectrum, which may be due to the fact that this galaxy sample is not strongly clustered. In this case, the redshift-space power spectrum would be even closer to a linear power spectrum. In any case, assuming a linear power spectrum will still permit physically accurate matter field inference in redshift space (Erdoğdu et al. 2004). For this reason, in the absence of more precise information on the galaxy power spectrum in redshift space, here we will assume a linear power spectrum, calculated according to the prescription provided by Eisenstein & Hu (1998, 1999). One should also bear in mind that the data themselves will govern the inference process. For this reason, power spectra measured from the Hamiltonian samples will only be defined partially by the a priori power-spectrum guess and mostly by the data. However, we defer a more careful treatment of all physical effects including a joint inference of density field and power spectrum to a future work.

It is clear that precise correction of these redshift-space effects requires knowledge about the peculiar velocities of all observed galaxies, which is usually not provided by galaxy redshift surveys. Therefore, precise correction of redshift-space distortions is very complicated and subject to ongoing research. In the linear regime, the theory behind the observed redshift-space distortions is well developed (Kaiser 1987; Hamilton 1998). However, in quasi-linear and non-linear regimes, we instead have to resort to making approximations or using fitting formulae based on numerical simulations (Percival & White 2009). The literature provides numerous approaches to alleviate these redshift-space distortions, particularly in power-spectrum estimation. Most of these approaches aim at restoring the correct power by deconvolution with a redshift-space convolution kernel which takes into account the random pair velocities of galaxies in collapsed objects (see e.g Peacock & Dodds 1994; Ballinger, Peacock & Heavens 1996; Hamilton 1998; Jing, Mo & Boerner 1998; Kang et al. 2002; Erdoğdu et al. 2004; Jing & Börner 2004; Scoccimarro 2004; Cabré & Gaztañaga 2009; Percival & White 2009). Such techniques have been adopted to correct Wiener density reconstructions by applying a redshift distortion operator to the final result, in order to restore the correct

power (Erdoğdu et al. 2004; Kitaura et al. 2009). However, it must be noted that this method does not account for the correction of phase information and therefore only corrects the two-point statistics of the recovered density field. Correcting also the phases of the density field will rather need non-linear approaches than simple deconvolution techniques.

Three-dimensional density inference hence requires redshift-space distortion corrections which also account for phase information and would be dependent on the density or gravitational potential. In the linear regime, it is possible to apply an inverse redshift-space operator which transforms the redshift-space density to its real-space counterpart (Nusser & Davis 1994; Taylor & Valentine 1999; D'Mellow & Taylor 2000). However, it does not account for the strongly non-linear regime which mostly generates the *Finger-of-God* effect. For this reason, Tegmark et al. (2004) proposed a *Finger-of-God* compression method. Here, they use a standard friends-of-friends algorithm to identify a cluster by taking into account different density thresholds, which set the linking length. They then measure the dispersion of galaxies about the cluster centre along the line of sight and in the transverse direction. If the radial dispersion exceeds the transverse dispersion, the cluster is compressed radially until the radial dispersion equals the transverse dispersion (Tegmark et al. 2004). However, it is not clear to what degree such a method would falsely isotropize filaments or under dense objects along the line of sight to spherical clusters. Such a method of isotropizing the density field, however, can also be applied in a post-processing step, by noting that a density threshold refers to a linking length in the friends-of-friends algorithm.

Nevertheless, the above correction methods mask the fact that redshift-space distortions introduce statistical uncertainties. Thus, unique recovery of the real-space density field is generally not possible. A full characterization of the joint uncertainties of the real-space density hence would require to carefully take into account the uncertainties introduced by redshift-space distortions or the lack of knowledge on peculiar velocities. This can be achieved by introducing a density-dependent peculiar velocity sampling scheme to our method, as proposed by Kitaura & Enßlin (2009). However, we defer sampling of the peculiar velocities, and therefore correction of redshift distortion effects, to a future work.

4 RESULTS

In this section, we describe the results obtained from the large-scale structure inference procedure.

4.1 Convergence test

HADES is a Markov chain Monte Carlo sampler and hence we have to test if the individual Hamiltonian samples really represent the log-normal Poissonian posterior. Convergence diagnostic of Markov chains is the subject of many discussions in the literature (see e.g. Heidelberger & Welch 1981; Gelman & Rubin 1992; Geweke 1992; Raftery & Lewis 1995; Cowles & Carlin 1996; Hanson 2001; Dunkley et al. 2005). However, here we apply the widely used Gelman & Rubin diagnostic, which is based on multiple simulated chains by comparing the variances within each chain and the variance between chains (Gelman & Rubin 1992). In particular, we calculate the potential scale reduction factor (PSRF; see Jasche & Kitaura 2010). A large PSRF indicates that the interchain variance is substantially greater than the intrachain variance, and longer chains are required. Once the PSRF approaches unity, one can conclude that each chain has reached the target distribution. We calculated the PSRF for each voxel in our calculation domain. The result for this test is presented in Fig. 1. It indicates convergence of the Markov chain. However, it can be seen that some regions of the domain converge faster than others. This is due to the fact that not all regions of the cubical volume have been observed equally. Regions which contain good observations converge faster, since there the probability distribution is narrower, while poorly or non-observed regions converge slower, since the space of possible solutions is larger. Also, note that the Gelman & Rubin diagnostic is generally a conservative test, and other tests might indicate convergence much earlier. However, this test clearly demonstrates that the quality and amount of observational data can have a strong impact on the convergence behaviour of the chain.

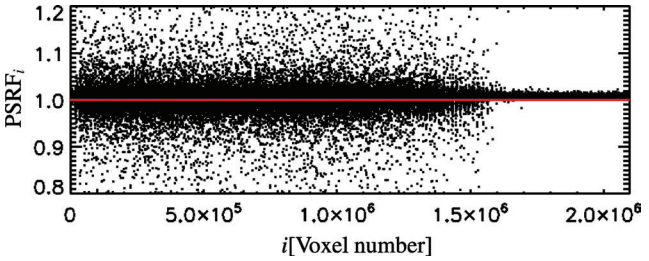

Figure 1. Results of the Gelman & Rubin convergence diagnostic. The PSRF indicates convergence. As can be seen, the Gelman & Rubin test converges faster in regions with good data.

4.2 Hamiltonian samples

Since the Markov chain converges, we can conclude that the individual samples are really samples from the large-scale structure posterior. At this point, it is important to insist that the Hamiltonian samples are not the result of a filtering procedure. A filter generally suppresses the signal in low-signal-to-noise ratio regions and therefore produces biased estimates for the physical density field. This is not the case for the individual Hamiltonian samples. Since they are random realizations of the lognormal Poissonian posterior, they are unbiased density fields in the sense that they possess correct physical power throughout the entire cubical volume. As an example, we present slices through an arbitrary density sample in Fig. 2. Already visually, one has the impression that the density field has equal power throughout the entire domain, even in the unobserved regions. This is because the Hamiltonian sampler non-linearly augments the poorly or unobserved regions with statistically correct information. Each density sample therefore is a proper physical density field, from which physical quantities can be derived. To demonstrate this, we measure the power spectra of some of these Hamiltonian samples. The result is presented in Fig. 3. As can be seen, the power spectra of the individual samples are very close to the assumed linear ΛCDM power spectrum. The deviations at large scales and small scales are due to the impact of the data. At small scales the deviation can be explained by redshift-space distortions, while at the largest scales cosmic variance is dominant. There is clearly no sign of artificial power loss due to the survey geometry. Another important issue to repeat at this point is that the aim of the Hamiltonian sampler is not to provide a single unique reconstruction of the underlying density field but rather to explore the space of possible full three-dimensional density fields, which are compatible with the observation. A unique reconstruction of the three-dimensional density field is generally impossible, particularly due to the statistical uncertainties in the observations. For

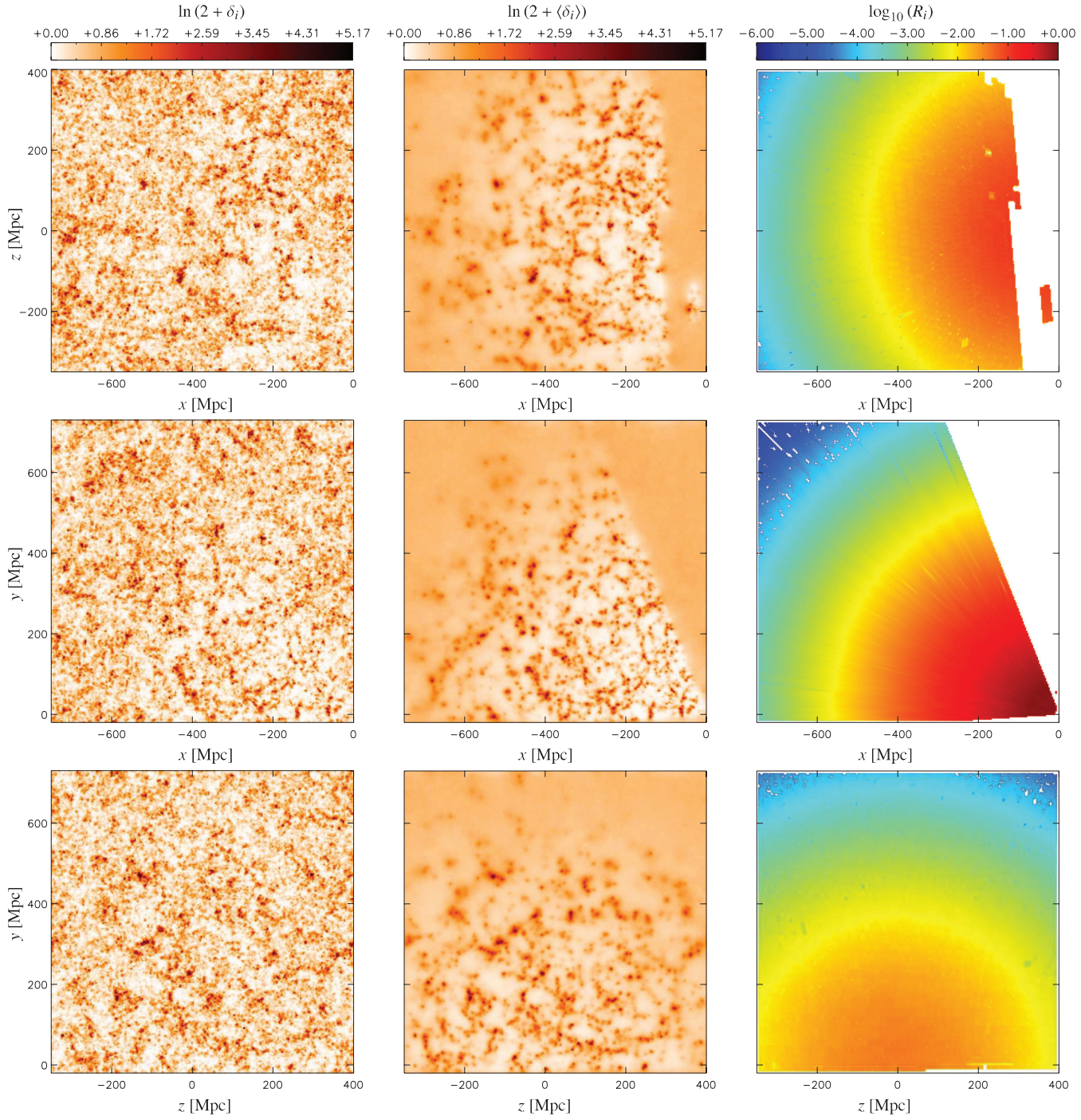

Figure 2. Three different slices from different sides through density fields. The left-hand panels show slices through one of the 40 000 density samples, middle panels depict the estimated ensemble mean and right-hand panels demonstrate the according slices through the three-dimensional response operator R_i. It can be seen that the density sample (left-hand panels) possesses equal power throughout the entire domain, even in the unobserved regions.

this reason, the Hamiltonian sampler provides us with a sampled representation of the cosmic density posterior distribution, which contains all the information on the three-dimensional density field, which can be extracted from the data and at the same time quantifies its uncertainties. Each individual Hamiltonian sample therefore represents a possible three-dimensional large-scale structure configuration, which is compatible with the observation. Once we have obtained this set of Hamiltonian samples, we can easily calculate statistical summaries such as the ensemble mean, of which we show some slices in the middle panels of Fig. 2. It can be seen that the ensemble mean nicely reflects the filamentary structure of the cosmic web. Further, in the unobserved regions the density fluctuations of the individual samples add up to cosmic mean density on average. This is of course an expected result since on average, we would expect mean density to be found in unobserved regions in the sky. Also note that in poorly observed regions, where the sampling of the density field by galaxies is bad due to selection effects, the statistical uncertainties are also high, allowing the individual density samples

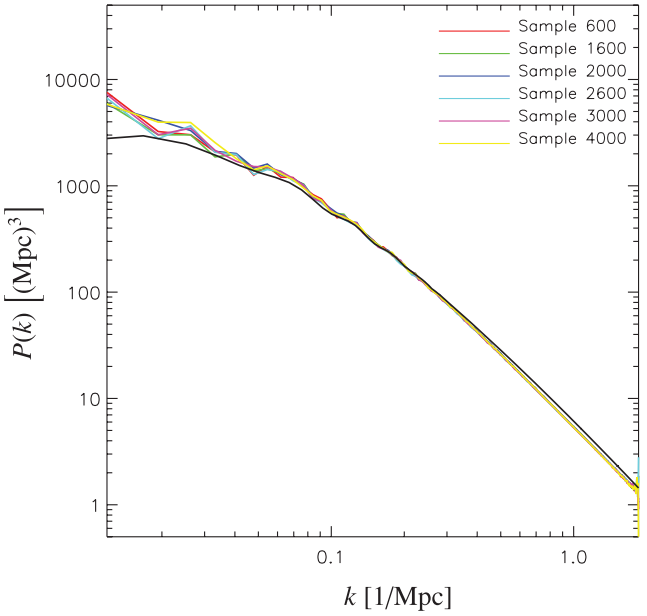

Figure 3. Power spectra of some Hamiltonian samples. The black curve corresponds to a linear ΛCDM power spectrum.

to deviate largely from the ensemble mean density field. This can be particularly observed in regions which are far away from the observer, where only small numbers of galaxies are observed. On the other hand, in regions with very good observations, the individual density samples only fluctuate little around the ensemble mean, reflecting the smaller uncertainty in these regions due to a larger signal-to-noise ratio. In this fashion, each individual density sample represents a random realization of the lognormal Poissonian posterior. Also note that since the individual samples are valid density field realizations, it is easy to derive meaningful physical quantities, such as the gravitational potential, cosmic flows or the tidal shear tensor as demonstrated in the remainder of this paper.

4.3 Ensemble mean and variance

Here we want to present the ensemble mean and variance for the set of 40 000 Hamiltonian samples, each consisting of 256^3 voxels. For comparison with a single density sample the middle panels of Fig. 2 show the according slices through the ensemble mean density field, which exhibits many interesting features. First, it renders remarkably well the filamentary structure of our cosmic neighbourhood. Many clusters, filaments and voids can clearly be seen by visual inspection. In the unobserved regions, the ensemble mean density amplitudes drop to the cosmic mean for the density contrast $\delta = 0$, just as required by construction. Structures close to the observer, at Cartesian coordinates (0, 0, 0), are more clearly visible than structures at larger distances. Especially, filaments and voids are less prominent at larger distances. This is due to the observational response operator R_i, which due to the radial selection function drops to very low values at large distances. Therefore, once a galaxy is detected far away from the observer, it must reside inside a large overdensity and hence inside a cluster. This expectation is clearly represented by the ensemble mean density field. Another interesting point to remark is that the borders to the unobserved regions are not very sharp. Some of the observed information is non-linearly propagated into the unobserved regions, since our method takes into account the correlation structure of the underlying signal. It can therefore be seen that some clusters and voids are interpolated further out into the unobserved regions. In comparison to the Wiener filter as previously applied to SDSS data by Kitaura et al. (2009), it seems that the Hamiltonian sampler is more conservative and less optimistic for the extrapolation of information into the unobserved region. This may be due to the fact that here we take into account the full Poissonian noise statistics rather than restricting the noise to a Gaussian approximation. Beside the ensemble mean, here we also calculate the ensemble variance per voxel, which is the diagonal of the full ensemble covariance matrix. Some slices through the ensemble mean, ensemble variance and the according slices through the observational response operator are presented in Fig. 4. Here the middle panels correspond to ensemble variance. At first glance, one can nicely see the Poissonian nature of the galaxy shot noise. High density peaks in the ensemble mean map correspond to high variance regions in the ensemble variance map, as expected for Poissonian noise. One can clearly see that the Hamiltonian sampler took into account the full three-dimensional noise structure of the galaxy distribution. Additionally, with a larger distance to the observer, the average variance increases, as is expected due to the radial selection function. It is also interesting to remark that some voids have been detected with quite low variance, hence with high confidence. Note, however, that although here we only plotted the diagonal of the density covariance matrix, the full non-diagonal covariance structure is completely encoded in the set of Hamiltonian samples and can be taken into account for future analysis. Also, note that the variance slices show high variances in regions where many galaxies have been observed. This is a key feature of the Poisson statistics, because the standard deviation is equal to the square root of the number of individual galaxies. That is, if there are N galaxies in each voxel, the mean is equal to N and the standard deviation is equal to $\sqrt{N}$. This makes the signal-to-noise ratio equal to $\sqrt{N}$ for such a homogeneous case. To emphasize the fact that regions which show high variances have also high signal-to-noise ratios, we calculate the density-to-variance ratio:

$$\omega_i = \frac{(1 + \langle \delta_i \rangle)}{\sqrt{\left\langle \delta_i^2 \right\rangle - \langle \delta_i \rangle^2}} \,. \tag{13}$$

The result of this calculation is presented in Fig. 5 for the case of the lower slices of Fig. 4. It clearly indicates high signal-to-noise ratios in high density regions. In addition, we also estimate the cumulative probabilities $\mathcal{P}\,(\delta_i \leq \delta_{\rm th})$ at 20 different density threshold values $\delta_{\rm th}$, for the density found at each voxel. These cumulative probabilities are estimated from the Hamiltonian samples by

$$\mathcal{P}\,(\delta_i \leq \delta_{\rm th}) = \frac{\sum_{n=1}^{N_{\rm samp}} \Theta(\delta_{\rm th} - \delta_i)}{N_{\rm samp}} \,, \tag{14}$$

where n labels the individual Hamiltonian samples, $N_{\rm samp}$ is the total number of samples and $\Theta(x)$ is the Heaviside function. These cumulative probabilities allow, for example, to estimate the median density at each voxel and can be useful when analysing galaxies in their cosmic environment as will be done in a following project. Some such cumulative probability distributions, chosen randomly, are shown in Fig. 6. As can be seen, the recovered density amplitudes extend over a large range, from small linear to very high non-linear values.

5 WEB CLASSIFICATION

In Section 1, we have already mentioned that the results presented in Section 4 are to be used for analysing galaxy properties in the

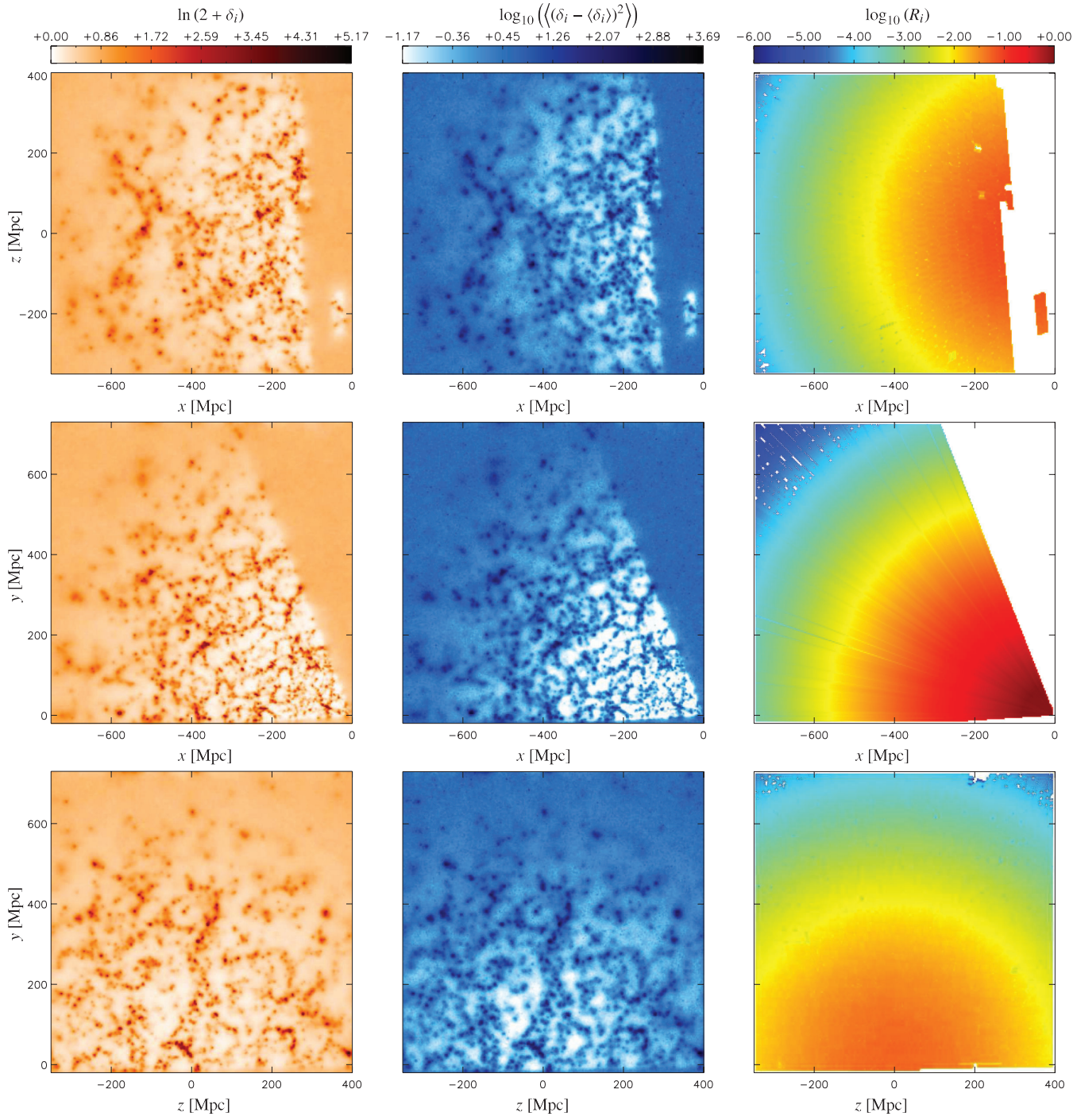

Figure 4. Three different slices from different sides through ensemble mean density (left-hand panels), ensemble variance (middle panels) and the three-dimensional response operator R_i (right-hand panels). Especially the variance plots demonstrate that the method accounted for the full Poissonian noise structure introduced by the galaxy sample. One can also see the correlation between high density regions and high variance regions, as expected for Poissonian noise.

large-scale environment in a future work. Such analyses also require the classification of the large-scale density field into different web-type objects. Therefore, in order to characterize the environment of our SDSS galaxy population, here we apply the dynamic web classification procedure, as described in Section 2.3, to the set of Hamiltonian samples. A similar analysis has been previously carried out by Lee & Erdogdu (2007) and Lee & Lee (2008) based on a Wiener mean density reconstruction of the Two-Micron All-Sky Survey (2MASS) redshift survey to study alignments of galaxy spins with the tidal field and the variation of a galaxy morphological type with environmental shear.

Here we will follow a similar procedure to classify each individual voxel of a given Hamiltonian sample into one of the four web types T_i, these types being T_1 = void, T_2 = sheet, T_3 = filament, T_4 = halo. To do so, we perform the following three steps for an individual Hamiltonian sample.

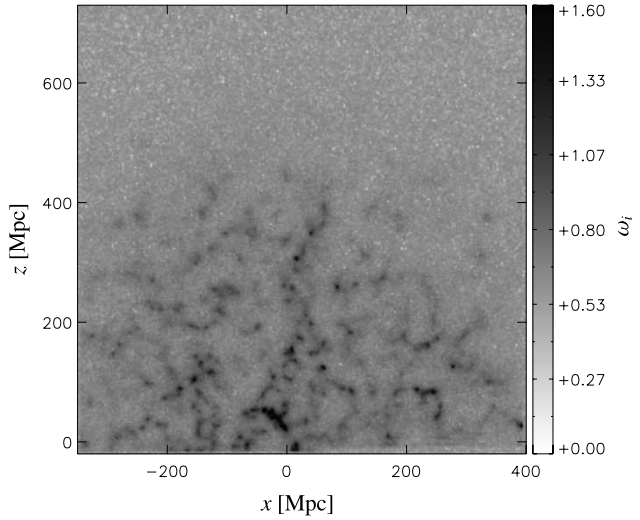

Figure 5. The plot shows the relative density to variance ratio ω_i. In comparison to the lower panels of Fig. 4, it indicates a high signal-to-noise ratio in regions of high density as expected for Poissonian noise.

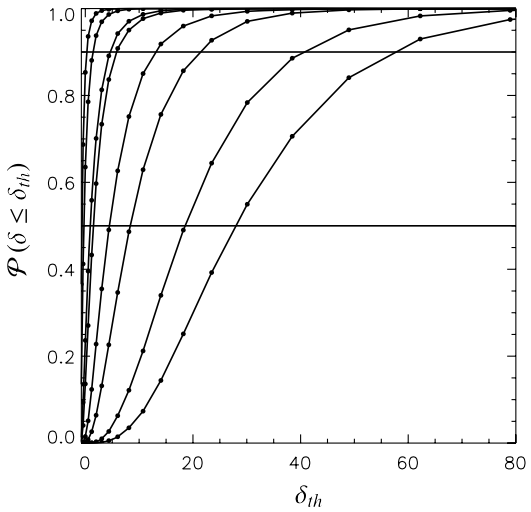

Figure 6. Cumulative probability distributions of the density at randomly chosen points in the volume. The cumulative probability distributions have been evaluated for 20 threshold values $\delta_{\rm th}$. The two horizontal lines indicate the $\mathcal{P}(\delta \leq \delta_{\rm th}) = 0.5$ and 0.9 thresholds, respectively.

(i) Solve equation (6) for the deformation tensor T_{ij} by means of fast Fourier transform techniques.

(ii) Solve the cubic characteristic equation for the three eigenvalues of the deformation tensor at each spatial position.

(iii) Apply the rules given in Table 1 to classify the web type at each spatial position for a given threshold value $\lambda_{\rm th}$.

The result of this procedure for the nth sample is then a unit 4-vector $\boldsymbol{T}^n(\boldsymbol{x}_k)$ at each voxel position $\boldsymbol{x}_k$. All of the entries of this 4-vector are zero except for one, which indicates the web type.

However, at this point it should be noted that the density fields have been obtained in redshift space, and hence particularly redshift-space distortions might change the morphological structure of the real-space density field especially in high density regions. Nevertheless, in the absence of a clear gauging of the threshold variable $\lambda_{\rm th}$ there still exists a degree of arbitrariness which may be more important for the clear definition of structure types. A thorough investigation of the impact of redshift-space distortion effects and their correction will be deferred to a future work. In the following, we will use the web-type classification as an example of simple non-linear and non-Gaussian error propagation.

Applying the method, as outlined above, to all Hamiltonian samples will yield a set of classification 4-vectors, which encodes the information and uncertainty of the observations. Additionally, as an intermediate result, we obtain the set of the three eigenvalues for each individual Hamiltonian sample. Slices through their ensemble mean estimates are presented in Fig. 7.

However, rather than summarizing the results in terms of mean and variance here we want to estimate the full cosmic web posterior. This is achieved by counting the relative frequencies for each web type at each individual spatial coordinate within the set of Hamiltonian samples. With these definitions, we yield the cosmic web posterior for each web type as

$$\mathcal{P}\left(T_i(\boldsymbol{x}_k)|\left\{N_k^{\rm g}\right\},\lambda_{\rm th}\right) = \frac{\sum_{n=1}^{N_{\rm samp}}\sum_{j=1}^{4}\delta^K_{T_i(\boldsymbol{x}_k)\,T_j^n(\boldsymbol{x}_k)}}{N_{\rm samp}}\,, \tag{15}$$

where n labels the individual Hamiltonian samples, $N_{\rm samp}$ is the total number of samples and δ^K_{ij} is the Kronecker delta. The cosmic web posterior incorporates all observational information and uncertainties and enables us to determine how well different structures can be classified with respect to observational uncertainties.

We evaluate the cosmic web posterior for four different values of $\lambda_{\rm th}$, with $\lambda_{\rm th} = 0.0,\ 0.2,\ 0.4,\ 1.0$. Slices through the cosmic web posteriors for the four different cases are presented in Fig. 8. It can be clearly seen that the properties of the survey geometry are represented by the four posterior distributions. While the web classification in the observed regions clearly follows the structure of the underlying density field, it obviously cannot provide a clear classification of unobserved regions. Also with increasing distance to the observer, the web classification becomes more and more uncertain. In this fashion, the cosmic web posterior renders the uncertainties introduced by the radial selection function and the resulting higher shot noise contribution at larger distances. The impact of the $\lambda_{\rm th}$ threshold can be observed when comparing the four cosmic web posteriors. In the case of $\lambda_{\rm th} = 0.0$ the cosmic web consists of many small isolated voids, which occupy only a small fraction of the total area of the slice. With increasing threshold $\lambda_{\rm th}$, voids become bigger and more connected until they completely dominate the cosmic web for the case $\lambda_{\rm th} = 1.0$. The opposite behaviour can be observed in the case of the halo posteriors, as the number of clearly detected haloes declines with increasing threshold $\lambda_{\rm th}$. Following Forero-Romero et al. (2009), we also calculate the volume occupied by each web type [volume filling fraction (VFF)] and the fraction of mass contained in such a volume [mass filling fraction (MFF)]. The results are presented in Fig. 9 and show the same behaviour as described in Forero-Romero et al. (2009). Fig. 9 supports the visual impression, gained by inspection of Fig. 8, that especially the VFF and MFF for voids strongly depend on the threshold value $\lambda_{\rm th}$. This shows that voids can serve as a sensitive monitor and an indicator of the cosmic web (Forero-Romero et al. 2009). Unfortunately, Forero-Romero et al. (2009) do not provide an explicit gauging of the $\lambda_{\rm th}$ values from simulations. Such a gauging and hence a clear definition of the different cosmic web types would be very valuable for these types of analysis.

Having now a representation of the web-type posterior, we can for example calculate the odds $O_i(\boldsymbol{x}_k)$ ratio given as

$$O_i(\boldsymbol{x}_k) = \frac{\mathcal{P}\left(T_i(\boldsymbol{x}_k)|\left\{N_k^{\rm g}\right\},\lambda_{\rm th}\right)}{1-\mathcal{P}\left(T_i(\boldsymbol{x}_k)|\left\{N_k^{\rm g}\right\},\lambda_{\rm th}\right)}\,\frac{1-\mathcal{P}\left(T_i(\boldsymbol{x}_k)\right)}{\mathcal{P}\left(T_i(\boldsymbol{x}_k)\right)}\,, \tag{16}$$

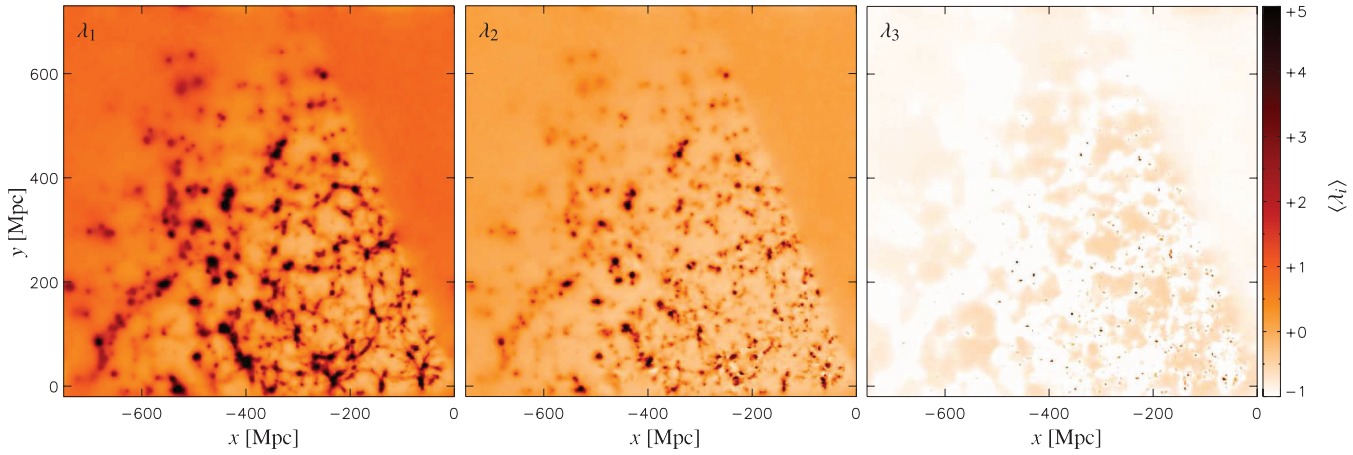

Figure 7. Ensemble mean of the eigenvalues of the deformation tensor.

which tells us how much a specific web type is favoured over all others. Here, the $\mathcal{P}(T_i(\boldsymbol{x}_k))$ can be obtained by averaging over all voxels in the volume. For example, this permits us to build a simple structure-type map $m(\boldsymbol{x}_k)$ which can be used for visual analyses as presented in the next section. Such a map can be defined as

$$m(\boldsymbol{x}_k) = \begin{cases} T_i(\boldsymbol{x}_k) & \text{for } O_i(\boldsymbol{x}_k) \geq O_{\rm th} \\ \text{undecided} & \text{else} \end{cases}, \qquad (17)$$

where $O_{\rm th}$ is an odds threshold usually chosen larger than unity.

6 GALAXY PROPERTIES VERSUS LARGE-SCALE STRUCTURE

In this section, we present a preliminary, but intuitive examination of the correlations between the large-scale environment of galaxies and their physical properties. Here we consider two properties of galaxies: stellar mass $M_\star$ and $g-r$ colour, and study how these are correlated with the overdensity δ of the large-scale environment and its type, which is one of the four web types classified as halo, filament, sheet and void. We will come back to this topic in a separate paper by considering more physical properties of galaxies and performing more careful and quantitative analyses.

Our results are shown in Figs 10 and 11 where we plot the galaxies in our sample with different stellar masses and $g-r$ colours, on top of a slice through the ensemble mean density field. In each figure, the four panels correspond to four $M_\star$ intervals as indicated. The galaxies falling into a given $M_\star$ range are plotted in the corresponding panel, with red (blue) galaxies being shown as red (blue) dots. Here we classify each galaxy into red or blue population using its $g-r$ colour and the luminosity-dependent divider as described in Li et al. (2006) (see their equation 7 and table 4). The observer on Earth is at the bottom right-hand corner of the slice where $x=0$ and $y=0$ Mpc. The density field with $z = 302.16 \pm 4.5$ Mpc is projected on to the $x-y$ plane and is repeated in every panel. In Fig. 10 the background density field is coded by the mean overdensity, $\ln(2+\langle\delta_i\rangle)$, averaged for each pixel over the z range probed and the 40 000 Hamiltonian samples. In Fig. 11, we present a structure-type map as defined in equation (17) by choosing an odds threshold of $O_{\rm th}=1.55$ and $\lambda_{\rm th}=1.0$. Each pixel of this map is colour-coded by the web type which is determined by our classification algorithm described above, with types of halo, filament, sheet and void being plotted in black, light grey, dark grey and white, respectively.

Qualitatively, the galaxies plotted in these figures appear to closely trace the underlying large-scale structure. This is not surprising because, by construction, the latter is reconstructed from the former. However, careful comparison of the different panels reveals a number of interesting trends. First, there exists a clear correlation between the galaxy mass and the large-scale environment, regardless of how the environment is quantified. More massive galaxies tend to reside in regions with higher densities and more halo-like structures. At the highest masses, almost all galaxies are confined within regions of high densities or those of halo and filament types. As $M_\star$ decreases, more and more galaxies are found in void-like regions. Secondly, at fixed stellar mass, the galaxy colour also appears to be correlated with the large-scale environment. Red galaxies trace the density field more closely than blue galaxies. At all masses, the distribution of blue galaxies is more extended across the different types of structures. At low masses, the blue population dominates the galaxies in void-like environment.

These trends are consistent with recent similar studies by Lee & Lee (2008) and Lee & Li (2008), which were based on much shallower galaxy samples (thus smaller volume), and also with the clustering analyses of Li et al. (2006). More work is needed in order to have more quantitative characterization of the relationships between galaxy properties and the large-scale environment, and thus more powerful constraints on galaxy formation models. These results, in turn, can be fed back to the large-scale structure inference and help to improve our cosmographical description of the Universe.

7 SUMMARY AND CONCLUSION

In this work, we present the first application of the non-linear, non-Gaussian Bayesian large-scale structure inference algorithm HADES to SDSS DR7 data.

HADES is a numerically efficient implementation of a Hamiltonian Markov chain sampler, which performs sampling in extremely high parameter spaces usually consisting of $\sim 10^7$ or more free parameters. In particular, HADES explores the lognormal Poissonian density posterior, which permits precision recovery of poorly sampled objects and density field inference deep into the non-linear regime (Jasche et al. 2010).

The large-scale structure inference was conducted on a cubic equidistant grid with a side length of 750 Mpc consisting of 256^3 voxels, yielding a grid resolution of about 3 Mpc. The large-scale

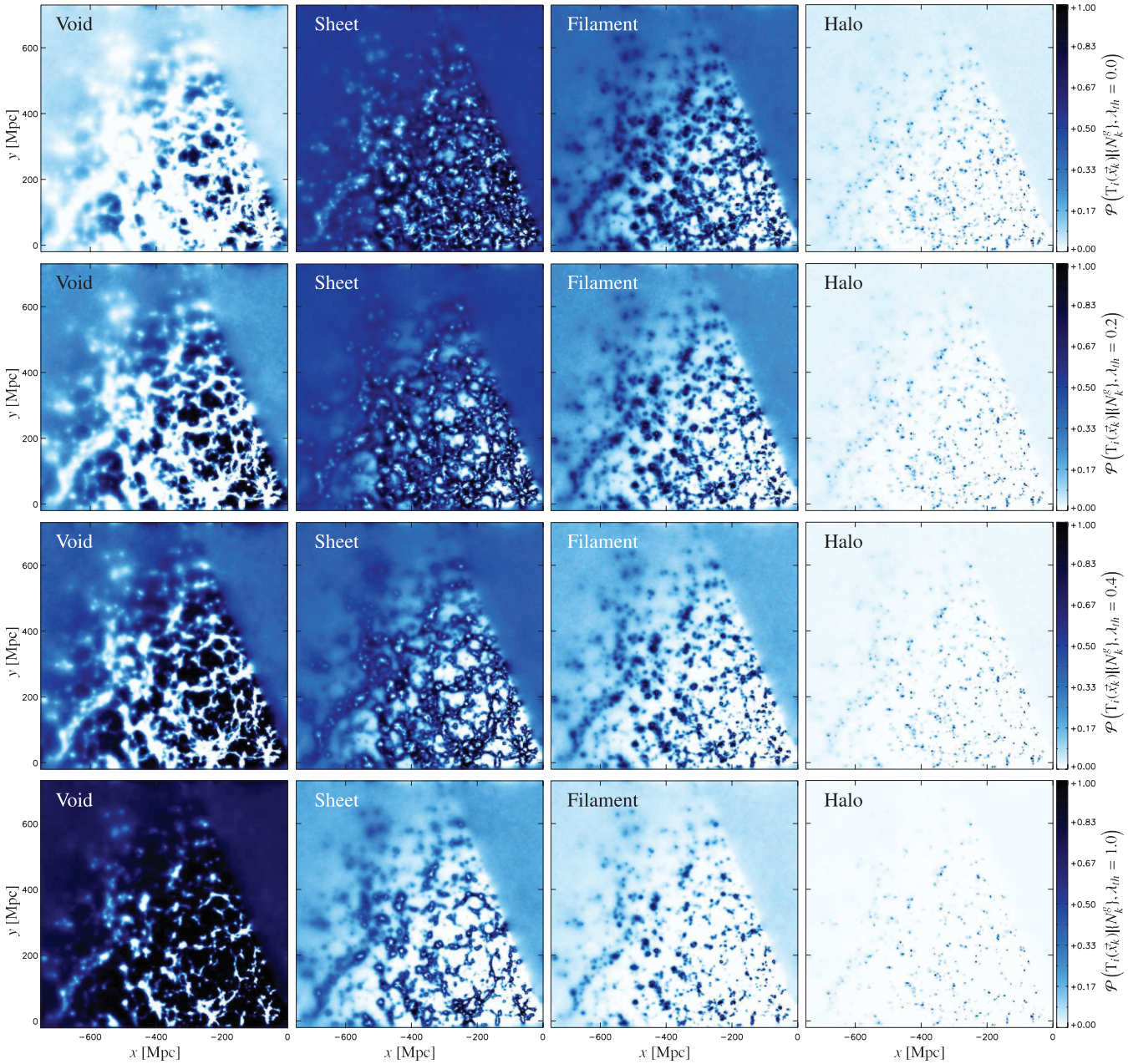

Figure 8. Slices through the cosmic web posterior for the threshold values $\lambda_{\rm th} = 0.0, 0.2, 0.4, 1.0$ (from top to bottom) for the four different web types. It is interesting to note that sliced sheets look filamentary, while filaments piercing the slice appear as dots.

structure inference procedure correctly accounts for the survey geometry, completeness and radial selection effects as well as for the correct treatment of Poissonian noise. The analysis yielded about 3TB of valuable scientific information in the form of full three-dimensional density samples of the lognormal Poissonian density posterior. This set of density samples is thus a sampled representation of the full non-Gaussian density posterior distribution and therefore encodes all observational systematics and statistical uncertainties. Hence, all uncertainties and systematics can seamlessly be propagated to any finally inferred quantity, by simply applying the according inference procedure to the set of samples. In this fashion, the results permit us to make precise and quantitative statements about the large-scale density field and any derived quantity.

We stress that our Hamiltonian samples are not the result of a filtering procedure. A filter generally suppresses the power of the signal in low signal-to-noise ratio regions and therefore does not yield a physical meaningful density, since it lacks power in poorly or unobserved regions. However, each Hamiltonian density sample represents a complete physical matter field realization conditional on the observations, in the sense that it possesses correct physical power throughout the entire volume. Visual inspection of these density samples has already shown a homogeneous distribution of power throughout the entire volume. This fact was emphasized by the demonstration of power spectra measured from these density samples, which show no sign of being affected by lack of power or artificial mode coupling nor do they show any sign of being affected by an adaptive smoothing kernel as would be expected for

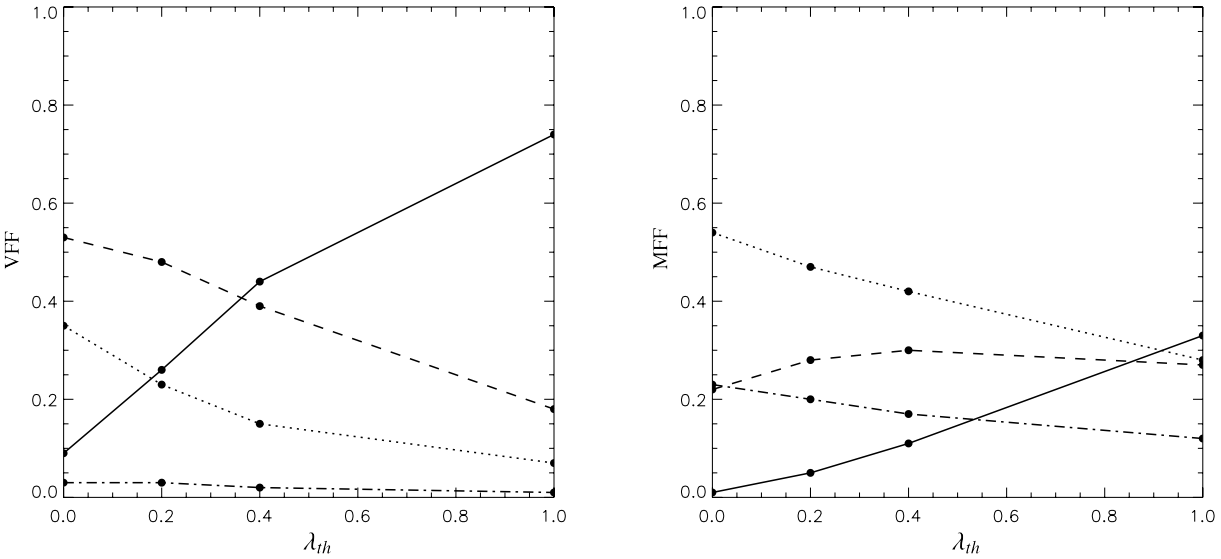

Figure 9. VFF and MFF as a function of $\lambda_{\rm th}$. Continuous lines denote voids, dashed lines sheets, dotted lines filaments and dot–dashed lines haloes. Especially the void VFF and MFF respond strongly to a change in $\lambda_{\rm th}$ making them a sensitive measure of the cosmic web (Forero-Romero et al. 2009).

filter applications. It should be noted that this fact marks the crucial difference of our method to previous filter-based density estimation procedures.

In Section 4.3, we estimated the ensemble mean and the according variance from the 40 000 density samples. The estimated ensemble mean nicely depicts the cosmic web consisting of filaments, voids and clusters extracted from the SDSS data. It is clear that the ensemble mean represents the mean estimated from the lognormal Poissonian posterior conditional on the SDSS data. Therefore, it encodes the observational uncertainties and systematics. This can be seen by the fact that the ensemble mean approaches cosmic mean density in poorly or unobserved regions. Further, we plotted the according variance, which demonstrates that the non-Gaussian behaviour and structure of the Poissonian shot noise were correctly taken into account in our analysis. Especially, the expected correlation between high mean density and high variance regions was clearly visible. We also estimated the cumulative probabilities for the density amplitude at each volume element and demonstrated that the recovered density fields truly cover the broad range from linear to non-linear density amplitudes.

To not only characterize the environment of our galaxy sample, but also to demonstrate the advantages of the Hamiltonian samples, we performed an example cosmic web-type classification in Section 5. In particular, we followed the dynamical cosmic web classification approach of Hahn et al. (2007) with the extensions proposed by Forero-Romero et al. (2009). This procedure involves the calculation of the cosmic deformation tensor and its eigenvalues. We demonstrated that this procedure can easily be applied to the set of samples, since they represent full physical matter field realizations. As a byproduct of this procedure, we estimated the ensemble mean for the three eigenvalues of the cosmic deformation tensor. Further, we classified the individual volume elements into one of the four different web types: void, sheet, filament and halo. The classification into four discrete web types enabled us to explicitly estimate the cosmic web posterior, which provides the probability of finding a specific web type at a given point in the volume conditional on the SDSS data. This result is especially appealing from a Bayesian point of view, since it emphasizes the fact that the result of a Bayesian method is a complete probability distribution rather than just a single estimate. Here we saw that especially voids are a sensitive measure for the cosmic web. Of course, it is possible to repeat the cosmic web classification in a similar manner to any other classification procedure.

In Section 6, we presented a preliminary examination of the correlation between the large-scale environment and physical properties of galaxies. In particular, we considered the stellar mass and $g - r$ colour of galaxies in relation to the density contrast δ. A qualitative analysis revealed that there exists correlation between these galaxy properties and the large-scale structure. In particular, massive galaxies are more likely to be found in massive structures, while low-mass galaxies reside in void-like structures. The plots also demonstrate the different clustering behaviour of red and blue galaxies. Also, note that these observed trends are consistent with previous works (Lee & Lee 2008; Lee & Li 2008; Li et al. 2006). However, more work is required in order to provide quantitative statements. This will be done in a forthcoming publication.

The results presented in this work will be valuable for many subsequent scientific analyses of the dependence of galaxy properties on their cosmic environment. In doing so, particularly the Hamiltonian samples allow for a more intuitive handling of observational data, since they can be understood as full matter field realizations or different multiverses consistent with our data of the Universe we live in. Besides providing quantitative characterizations of the large-scale structure, the results also give us an intuitive understanding of the three-dimensional matter distribution in our cosmic neighbourhood. We intend to make our data publically available to the community.

Future applications will also take into account non-linear bias models and peculiar velocity sampling procedures, to provide even more accurate density analyses and to accurately account for redshift-space distortions.

We hope that this work demonstrates the potential of Bayesian large-scale structure inference and its contribution to current and future precision analyses of our Universe.

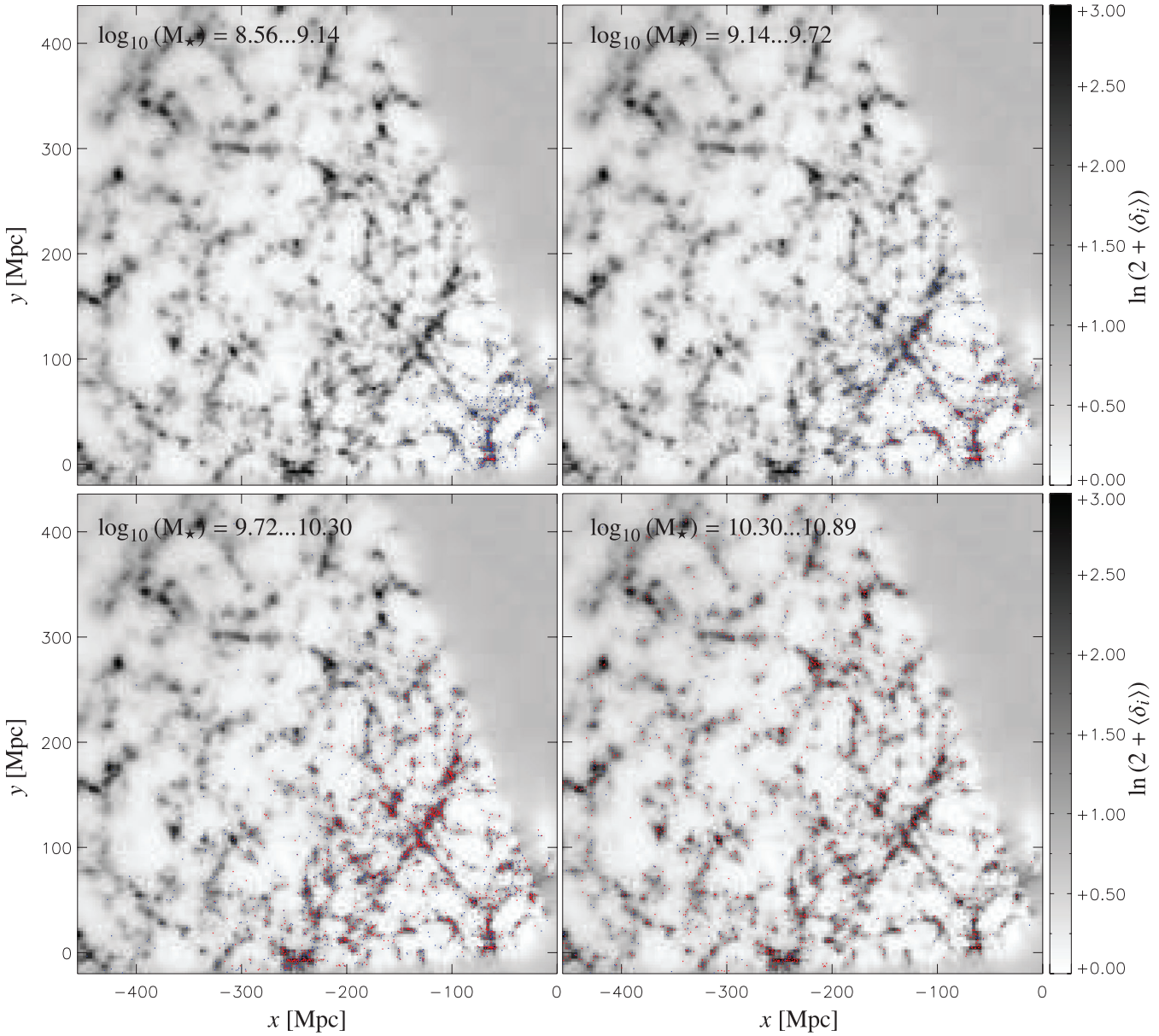

Figure 10. SDSS galaxies overplotted on the ensemble mean density field. The blue and red dots denote blue and red galaxies, respectively, and the different panels depict galaxies in different stellar mass $M_\star$ bins.

ACKNOWLEDGMENTS

We would like to thank Ofer Lahav and Benjamin D. Wandelt for suggesting us to use the lognormal Poissonian posterior for large-scale structure inference. We also thank Simon D. M. White for encouraging discussions. Particular thanks also to Rainer Moll and Björn Malte Schäfer for useful discussions and support with many valuable numerical gadgets. The authors also thank Benton R. Metcalf for many interesting discussions and comments on this project and Andreas Faltenbacher for suggesting us to estimate the cosmic deformation tensor. Special thanks also to María Ángeles Bazarra Castro for helpful assistance during the course of this project. Further, we thank the 'Transregional Collaborative Research Centre TRR 33 – The Dark Universe' for the support of this work.

Funding for the SDSS and SDSS-II has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, the U.S. Department of Energy, the National Aeronautics and Space Administration, the Japanese Monbukagakusho, the Max Planck Society and the Higher Education Funding Council for England. The SDSS web site is http://www.sdss.org/.

The SDSS is managed by the Astrophysical Research Consortium for the Participating Institutions. The Participating Institutions are the American Museum of Natural History, Astrophysical Institute Potsdam, University of Basel, University of Cambridge, Case Western Reserve University, University of Chicago, Drexel University, Fermilab, the Institute for Advanced Study, the Japan Participation Group, Johns Hopkins University, the Joint Institute for Nuclear Astrophysics, the Kavli Institute for Particle Astrophysics and Cosmology, the Korean Scientist Group, the Chinese Academy of Sciences (LAMOST), Los Alamos National Laboratory, the Max-Planck-Institute for Astronomy (MPIA), the Max-Planck-Institute for Astrophysics (MPA), New Mexico State University, Ohio State University, University of Pittsburgh,

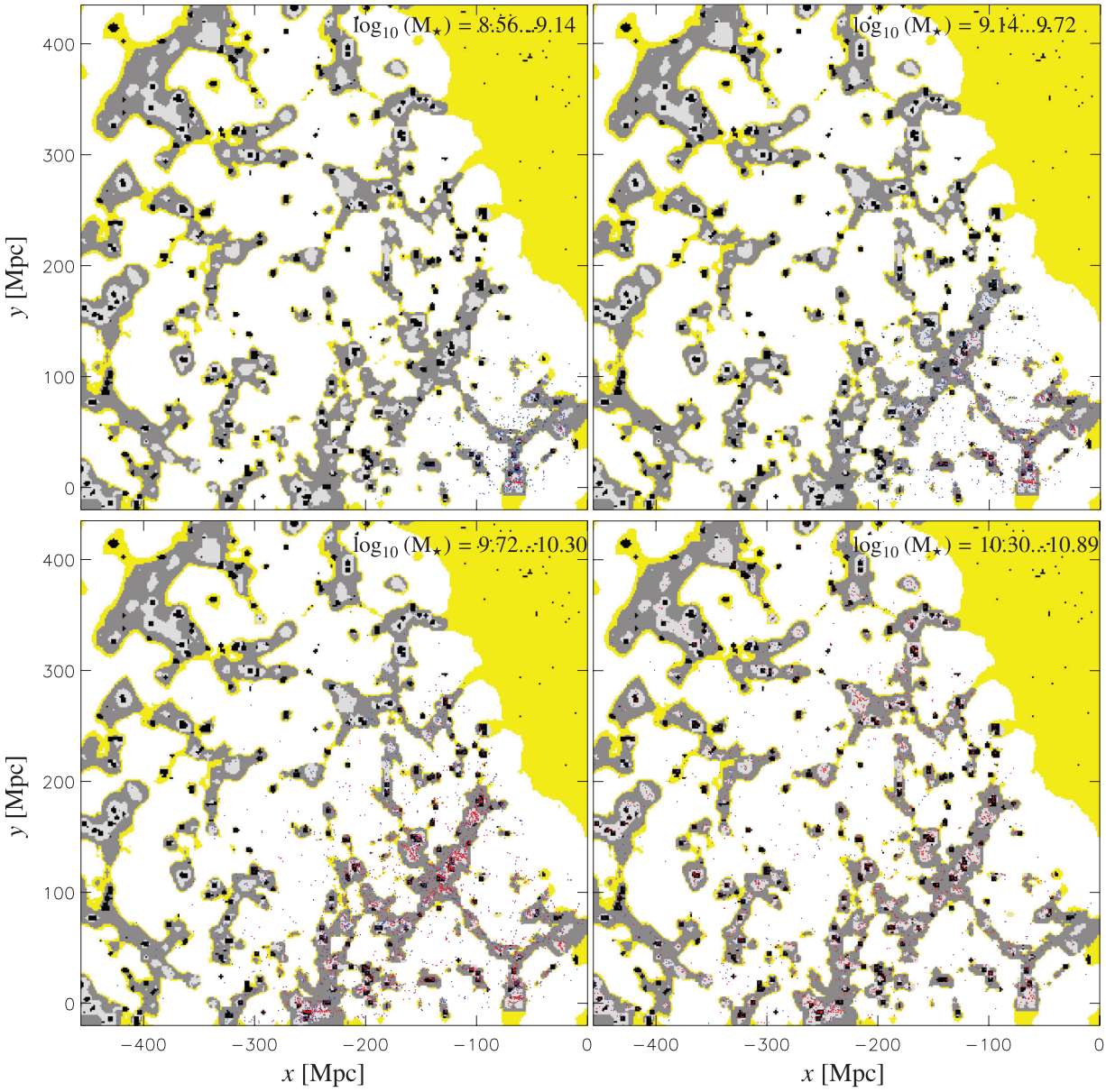

Figure 11. Same as Fig. 10, but here the galaxies are overplotted on a structure-type map as defined in Section 5. The colour coding denotes the web type: halo (black), filament (light grey), sheet (dark grey) and void (white). Regions, which are marked as undecided according to our criteria, equation (17) with $O_{\rm th} = 1.55$, are coloured yellow.

University of Portsmouth, Princeton University, the United States Naval Observatory and the University of Washington.

REFERENCES

Abazajian K. N. et al., 2009, ApJS, 182, 543
Alcock C., Paczyński B., 1979, Nat, 281, 358
Aragón-Calvo M. A., Jones B. J. T., van de Weygaert R., van der Hulst J. M., 2007, A&A, 474, 315
Ballinger W. E., Peacock J. A., Heavens A. F., 1996, MNRAS, 282, 877
Bernardi M., Nichol R. C., Sheth R. K., Miller C. J., Brinkmann J., 2006, AJ, 131, 1288
Bistolas V., Hoffman Y., 1998, ApJ, 492, 439
Blanton M. R., Roweis S., 2007, AJ, 133, 734
Blanton M. R. et al., 2003a, AJ, 125, 2348
Blanton M. R., Lin H., Lupton R. H., Maley F. M., Young N., Zehavi I., Loveday J., 2003b, AJ, 125, 2276
Blanton M. R., Eisenstein D., Hogg D. W., Schlegel D. J., Brinkmann J., 2005, ApJ, 629, 143
Cabré A., Gaztañaga E., 2009, MNRAS, 396, 1119
Choi Y., Park C., Vogeley M. S., 2007, ApJ, 658, 884
Colberg J. M., Sheth R. K., Diaferio A., Gao L., Yoshida N., 2005, MNRAS, 360, 216
Colberg J. M. et al., 2008, MNRAS, 387, 933
Coles P., Jones B., 1991, MNRAS, 248, 1
Cowles M. K., Carlin B. P., 1996, J. American Statistical Association, 91, 883
Davis M., Peebles P. J. E., 1983, ApJ, 267, 465
D'Mellow K. J., Taylor A. N., 2000, in Courteau S., Willick J., eds, ASP Conf. Ser. Vol. 201, Cosmic Flows Workshop. Astron. Soc. Pac., San Francisco, p. 298
Dressler A., 1980, ApJ, 236, 351

Duane S., Kennedy A. D., Pendleton B. J., Roweth D., 1987, Phys. Lett. B, 195, 216
Dunkley J., Bucher M., Ferreira P. G., Moodley K., Skordis C., 2005, MNRAS, 356, 925
Ebeling H., Wiedenmann G., 1993, Phys. Rev. E, 47, 704
Eisenstein D. J., Hu W., 1998, ApJ, 496, 605
Eisenstein D. J., Hu W., 1999, ApJ, 511, 5
Enßlin T. A., Frommert M., Kitaura F. S., 2009, Phys. Rev. D, 80, 5005
Erdoğdu P. et al., 2004, MNRAS, 352, 939
Erdoğdu P. et al., 2006, MNRAS, 373, 45
Fisher K. B., Lahav O., Hoffman Y., Lynden-Bell D., Zaroubi S., 1995, MNRAS, 272, 885
Forero-Romero J. E., Hoffman Y., Gottlöber S., Klypin A., Yepes G., 2009, MNRAS, 396, 1815
Frommert M., Enßlin T. A., Kitaura F. S., 2008, MNRAS, 391, 1315
Gaztanaga E., Yokoyama J., 1993, ApJ, 403, 450
Gelman A., Rubin D., 1992, Statistical Sci., 7, 457
Geweke J., 1992, in Bernardo J. M., Berger J., David A. P., Smith A. F. M., eds, Bayesian Statistics. Oxford Univ. Press, Oxford, p. 169
Gómez P. L. et al., 2003, ApJ, 584, 210
Goto T., Yamauchi C., Fujita Y., Okamura S., Sekiguchi M., Smail I., Bernardi M., Gomez P. L., 2003, MNRAS, 346, 601
Hahn O., Porciani C., Carollo C. M., Dekel A., 2007, MNRAS, 375, 489
Hamann J., Hannestad S., Melchiorri A., Wong Y. Y. Y., 2008, J. Cosmology Astropart. Phys., 7, 17
Hamilton A. J. S., 1998, Astrophysics and Space Science Library 231, 185
Hanson K. M., 2001, in Sonka M., Hanson K. M., eds, Proc. SPIE Conf. Ser. Vol. 4322, Markov Chain Monte Carlo Posterior Sampling with the Hamiltonian Method. SPIE, Bellingham, p. 456
Heidelberger P., Welch P. D., 1981, Commun. ACM, 24, 233
Hockney R. W., Eastwood J. W., 1988, Computer Simulation using Particles. Taylor & Francis, Inc., Bristol, PA
Hoffman Y., 1994, in Balkowski C., Kraan-Korteweg R. C., eds, ASP Conf. Ser. Vol. 67, Unveiling Large-Scale Structures Behind the Milky Way. Astron. Soc. Pac., San Francisco, p. 185
Hubble E., 1934, ApJ, 79, 8
Jasche J., Kitaura F. S., 2010, MNRAS, 407, 29
Jasche J., Kitaura F. S., Wandelt B. D., Enßlin T. A., 2010, MNRAS, 406, 60
Jing Y. P., Börner G., 2004, ApJ, 617, 782
Jing Y. P., Mo H. J., Boerner G., 1998, ApJ, 494, 1
Kaiser N., 1987, MNRAS, 227, 1
Kang X., Jing Y. P., Mo H. J., Börner G., 2002, MNRAS, 336, 892
Kayo I., Taruya A., Suto Y., 2001, ApJ, 561, 22
Kitaura F. S., Enßlin T. A., 2008, MNRAS, 389, 497
Kitaura F. S., Jasche J., Li C., Enßlin T. A., Metcalf R. B., Wandelt B. D., Lemson G., White S. D. M., 2009, MNRAS, 400, L183
Kitaura F. S., Jasche J., Metcalf R. B., 2010, MNRAS, 403, 589
Klypin A., Hoffman Y., Kravtsov A. V., Gottlöber S., 2003, ApJ, 596, 19
Kuehn F., Ryden B. S., 2005, ApJ, 634, 1032
Lahav O., 1994, in Balkowski C., Kraan-Korteweg R. C., eds, ASP Conf. Ser. Vol. 67, Unveiling Large-Scale Structures Behind the Milky Way. Astron. Soc. Pac., San Francisco, p. 171
Lahav O., Fisher K. B., Hoffman Y., Scharf C. A., Zaroubi S., 1994, ApJ, 423, L93
Layzer D., 1956, AJ, 61, 383
Lee J., Erdogdu P., 2007, ApJ, 671, 1248
Lee J., Lee B., 2008, ApJ, 688, 78
Lee J., Li C., 2008, preprint (arXiv:0803.1759)
Lemson G., Kauffmann G., 1999, MNRAS, 302, 111
Lewis I. et al., 2002, MNRAS, 334, 673
Li C., Kauffmann G., Jing Y. P., White S. D. M., Börner G., Cheng F. Z., 2006, MNRAS, 368, 21
Libeskind N. I., Yepes G., Knebe A., Gottloeber S., Hoffman Y., Knollman S. R., 2010, MNRAS, 401, L889
Magira H., Jing Y. P., Suto Y., 2000, ApJ, 528, 30
Martínez V., Saar E., 2002, Statistics of the Galaxy Distribution. Chapman & Hall/CRC, Boca Raton, FL
Martinez-Vaquero L. A., Yepes G., Hoffman Y., Gottlöber S., Sivan M., 2009, MNRAS, 397, 2070
Matsubara T., Suto Y., 1996, ApJ, 470, L1
Neal R. M., 1993, Technical Report CRG-TR-93-1, Univ. Toronto
Neal R. M., 1996, Lecture Notes in Statistics, Bayesian Learning for Neural Networks, 1st edn. Springer, Berlin
Novikov D., Colombi S., Doré O., 2006, MNRAS, 366, 1201
Nusser A., Davis M., 1994, ApJ, 421, L1
Park C., Choi Y., Vogeley M. S., Gott J. R. I., Blanton M. R., 2007, ApJ, 658, 898
Peacock J. A., Dodds S. J., 1994, MNRAS, 267, 1020
Peacock J. A., Smith R. E., 2000, MNRAS, 318, 1144
Peebles P., 1980, Large Scale Structure of the Universe. Princeton Univ. Press, Princeton, NJ
Percival W. J., White M., 2009, MNRAS, 393, 297
Popowski P. A., Weinberg D. H., Ryden B. S., Osmer P. S., 1998, ApJ, 498, 11
Postman M., Geller M. J., 1984, ApJ, 281, 95
Raftery A. E., Lewis S. M., 1995, in Gilks W. R., Spiegelhalter D. J., eds, Practical Markov Chain Monte Carlo. Chapman & Hall, Boca Raton, FL, p. 115
Rojas R. R., Vogeley M. S., Hoyle F., Brinkmann J., 2005, ApJ, 624, 571
Saunders W., Ballinger W. E., 2000, in Kraan-Korteweg R. C., Henning P. A., Andernach H., eds, ASP Conf. Ser. Vol. 218, Mapping the Hidden Universe: The Universe Behind the Milky Way. Astron. Soc. Pac., San Francisco, p. 181
Saunders W. et al., 2000, in Kraan-Korteweg R. C., Henning P. A., Andernach H., eds, ASP Conf. Ser. Vol. 218, Mapping the Hidden Universe: The Universe Behind the Milky Way. Astron. Soc. Pac., San Francisco, p. 141
Scoccimarro R., 2004, Phys. Rev. D, 70, 083007
Seljak U., 2000, MNRAS, 318, 203
Sheth R. K., 1995, MNRAS, 277, 933
Smith R. E. et al., 2003, MNRAS, 341, 1311
Spergel D. N. et al., 2007, ApJS, 170, 377
Taylor A., Valentine H., 1999, MNRAS, 306, 491
Tegmark M. et al., 2004, Phys. Rev. D, 69
Tegmark M. et al., 2006, Phys. Rev. D, 74, 123507
van de Weygaert R., Schaap W., 2001, in Banday A. J., Zaroubi S., Bartelmann M., eds, Mining the Sky, Proc. MPA/ESO/MPE Workshop. Springer, Berlin, p. 268
Webster M., Lahav O., Fisher K., 1997, MNRAS, 287, 425
Whitmore B. C., Gilmore D. M., Jones C., 1993, ApJ, 407, 489
York D. G. et al., 2000, AJ, 120, 1579
Zaninetti L., 1995, A&AS, 109, 71
Zaroubi S., Hoffman Y., Fisher K. B., Lahav O., 1995, ApJ, 449, 446
Zaroubi S., Hoffman Y., Dekel A., 1999, ApJ, 520, 413

This paper has been typeset from a TEX/LATEX file prepared by the author.